The Future of Outcomes Measurement:
Item Banking, Tailored Short-Forms, and Computerized-Adaptive Assessment

Draft 1 – May 5, 2004

David Cella, Richard Gershon and Jin-Shei Lai

Center on Outcomes, Research and Education
Evanston Northwestern Healthcare and
Northwestern University Medical School

Prologue

The title of this session is presumptuous. It presumes the current way of conducting patient-reported outcomes (PROs), using instruments of fixed length in their entirety, will be replaced with a new approach. This new approach begins with clear definitions of important outcomes, and harnesses large, well-studied pools, or "banks" of PRO items. These items can be selected from the bank to form customized short scales, or can be administered in a sequence and length determined by a computer programmed for precision and clinical relevance. Is this a better way? Probably. Will it catch on? Possibly. What will help it catch on? Cooperation and some faith. Not the faith sometimes referred to by critics who liken item response theory modeling to religious dogma, but faith that despite imperfections in results and differences in approach, the real prize in store for the next 5-10 years is a common definition and understanding of human symptoms and functional problems such as fatigue, pain, depression, mobility, social function, sensory function, and many other health concepts that we can only measure by asking people about them. The support of the NIH, as witnessed in the hosting of this conference, the contracting with experts to build item banks and tying IRT to the NIH Roadmap Initiative through its commitment to PROMIS, is a big step in that direction. Consumers will come along if the producers agree on the basic quality of the products. Producers, in this case PRO measurement experts, must resist the temptation to elevate their own work and ideas at the expense of others with the same goal. If we really want this to work, it can. It has in other areas, like education and high stakes testing such as for professional licensure. These successes offer direction.

Perhaps because we are from Chicago, "The City that Works," our approach to item banking and computerized adaptive testing (CAT) is practical: As focused on application as it is on science or theory. From a practical perspective, we frequently must decide whether to re-write and retest an item, add more items to fill gaps (often at the ceiling of the measure), re-test a bank after some modifications, or split up a bank into units that are more unidimensional, yet less clinically relevant or complete. These decisions are not easy, and yet we find they are also rarely unforgiving. To borrow from an old phrase: There is more than one way to spin a CAT. (apologies for the bad pun) We encourage people to build practical tools that indeed spin out short form measures and CAT administrations from common banks, and to further our understanding of these banks with various clinical populations and ages, so that with time the scores that emerge from these many activities begin to have not only a common metric and range, but a shared meaning and understanding across users. We believe this can be the future, but recognize it is not the present. In this brief paper, we discuss our approach to item banking and its byproducts, offer a demonstration of CAT, describe testing options, discuss some work of others, and discuss models for long term sustainability of an IRT approach to PRO assessment. Some barriers to success we can anticipate are limitations in the methods themselves, controversies and disagreements across approaches, and end-user reluctance to move away from what is finally become comfortable in many settings.

Introduction

Advances in measurement using item response theory (IRT) and advances in computer technology make it possible to develop, maintain and improve item banks that can advance health status measurement. Item banks enable item comparison and selection as well as Computerized Adaptive Testing (CAT) tools for tailored individual assessment without loss of scale precision or content validity. Valid, generalizable item banks and CAT tools can stimulate and standardize clinical research across dealing with patient-reported outcomes (PROs). They also may assist individual clinical practitioners to assess patient response to interventions and modify treatment plans.

It is currently challenging to compare results across different studies where different measurement tools were used, as these instruments had non-comparable or non-combinable scores. New computer technologies and advances in measurement make it possible to: 1) develop, maintain and improve item banks; 2) compare items and conduct statistical modeling of responses; and 3) create CATs that allows item subsets to be tailored to the individual without loss of scale precision or content validity. Such a measurement system has great potential in clinical practice to rapidly and reliably assess response to interventions and to inform treatment modifications.

Overview of Item Banks

An item bank contains survey questions representative of a common theme or trait (such as pain or emotional well-being). When bank items are calibrated using IRT, each item is representative of the entire trait and has a characteristic precision in measuring that trait across its continuum. In contrast, classical test theory typically requires that an entire test be administered to appropriately represent the trait being measured. Bank items can be pre-selected from across the trait spectrum to produce a static instrument of any length, or to enable computerized adaptive testing (CAT). The scores produced by any of the instruments created from the calibrated bank are pre-validated on a single scale and are comparable regardless of the instrument used.

An item bank can be constructed to represent a trait that is generic (or common) across many diseases. Yet this bank can be employed to measure generic concepts in a way that is uniquely targeted to a given disease. For example, one could apply different filters to the same pain item bank such that diabetic patients received questions more targeted to neuropathic pain, whereas arthritis patients were administered more musculoskeletal pain questions. Items are typically only used in an item bank if the way patients respond to those items does not differ on the basis of diagnosis, language version (e.g. Spanish or English) or ethnicity (eg., Euro-American versus African American). Targeted (disease specific or treatment specific) scales represent traits that are not common across the oncology spectrum, but are otherwise important to defining HRQL for a particular disease. For example, treatment effects for prostate cancer may have little in common with treatment effects for breast cancer, but are important for understanding the HRQL of a person with prostate cancer. Items in targeted scales can be calibrated separately for each disease.

Figure 1 depicts some advantages of using IRT calibrated item banks compared to conventional instruments. A flexible, comprehensive array of clinically-relevant questions is categorized into a finite set of generic banks and additional targeted scales.

This is an important, novel departure from the conventional generic questionnaire plus targeted subscales now familiar to the clinical and HRQL research community. In the conventional approach, one generic instrument, perhaps 30-50 questions in length, assesses several core concepts important to patients across a spectrum of conditions. These concepts may include pain, fatigue, physical functioning, etc. Each concept is measured with a short set of questions scaled to produce a range of scores that can then be compared across patients with all conditions. Added to this static core instrument is then a set of condition-specific questions that are relevant to a subset (one or more) of the full list of targeted conditions. All patients across all conditions are administered the same generic questions; there is usually no discretion allowed the individual researcher. To claim assessment validity, all questions must be administered.

In contrast, using IRT, assessment of patients may be accomplished with an instrument even shorter than the static generic instrument. This is made possible by emphasizing and studying the individual item performance in relation to performance of other items in the set of questions representing a given concept being measured. In practice, when the IRT measurement model fits the item bank data, *one can select any subset of questions in that bank and derive a common standardized score*.

Figure 1 illustrates how three different hypothetical clinician/researchers, studying three different diseases, can access the same generic item bank and select unique short forms of varying length and clinical content, and produce a score for each person across the three trials that is on the same metric. IRT allows comparisons of patients and items across multiple instruments through a mechanism of equating the instruments along a common, interval scale, measurement continuum. The interval scale nature of the metric comfortably allows for parametric statistics to be applied to trial data, offering more power in the statistical test, and perhaps having a beneficial effect on sample size requirements for trials in which the sample size is driven by the HRQL endpoint.

Computerized Adaptive Testing (CAT)

Static scales in today's generic and targeted instruments are almost universally too coarse for individual classification and diagnosis. IRT item banks permit a degree of precision in assessing the individual person.{Gershon, 2003 9601 /id} CAT is a specific type of computer based testing that only asks questions that will provide a maximum amount of information. Using IRT measurement models, item selection is guided by an individual's response to previously administered questions from a large item bank. The respondent need only answer a small number of items to obtain a measure that accurately estimates what would have been obtained had the entire set of items been administered. Recently some have begun to apply CAT to PRO assessment.{McHorney, 2000 9172 /id;Ware, 2000 9039 /id} CAT is scored in real-time and results may be presented in graphic and/or written reports immediately to the physician and patient enabling them to have a focused discussion regarding treatment options and care planning.{Cella, 2000 210 /id;Wolfe, 1999 7633 /id} Patients have reported that such discussions improve communication with providers. These discussions help patients feel better understood by their physicians{Detmar, 2002 9234 /id;Velikova, 2003 9603 /id} and may encourage better care.{Jacobsen, 2002 1830 /id} Overall CAT offers several advantages including: 1) speed of assessment; 2) fewer items for the same level of precision; 3) mechanism for completing routine

assessments; 4) immediate data entry; 5) easier scoring and interpretation; 6) comparison of scores across time; 7) immediate presentation of results in “real-time”.

The logic behind CAT can be applied to surveys with polytomous (e.g., Likert-type) items. For example, if a person’s true level of pain is 75 on a scale of 0 to 100, the initial 5-choice rating scale item typically consist of options which span the entire range of the trait, perhsps corresponding to persons with pain levels of 20, 40, 60, 80 and 100. If the person selects the fourth option, our initial estimate of their level of pain is 80. A second item is then selected with options corresponding to pain levels of 60, 70, 80, 90 and 100. The person selects the option corresponding to 70, resulting in a new pain estimate of 75. A third item is then presented with options corresponding to pain levels of 65, 70, 75, 80 and 85. Items continue to be administered until a desired level of precision is obtained.

How we Create Item Banks

Our framework for item bank development guides our practice and tracks status (see Table 1). Literature review and input from clinicians and patients are used to select domains (step 1). We then survey the availability of relevant existing datasets (step 2). We then identify common items to serve as the linking items across datasets, and common rating scales (step 3). In cases where there are available datasets with sufficient sample size (minimum 200, preferably 500), we use a common item linking strategy to expedite the process and inform subsequent bank development. We examine unidimensionality and item fit, and then calibrate items on the measurement continuum (step 4). The item hierarchy is then examined, providing information regarding potential statistical construct deficiency, i.e., gaps in the distribution (step 5). Clinical construct deficiency is also identified. New items are written or acquired from existing questionnaires to eliminate construct deficiencies (step 6). For the banks that do not have existing datasets, we skip steps 3-5. Prior to field testing, content validity issues are addressed by obtaining clinical input to examine content relevance and representativeness (step 7). For field testing (step 8), items are then programmed for Computer Based Testing (CBT) and administered in clinical settings. The collected data are analyzed (step 9) to determine unidimensionality, item fit and item locations on the

Table 1. Item Bank Development Framework

Determining appropriate banks for development	
1. Determine domains to be covered in bank	a. Literature review
	b. Clinical input (incl.patients)
2. Determine availability of relevant data source(s)	If yes, go to step 3
	If no, go to step 6
Develop the initial item bank	
3. Identify common items and rating scales	
4. Data analysis	a. Examine dimensionality
	b. Examine item fit
	c. Calibrate items on the continuum
5. Examine construct deficiency	a. Statistical deficiency (gaps)
	b. Clinical deficiency (gaps)
Developing an operational item bank	
6. Acquire or write new item(s) with clinical input	
7. Content validation	
8. Field testing	a. CBT programming
	b. Data collection
9. Data analysis	a. Examine dimensionality
	b. Examine item fit
	c. Calibrate items on the continuum
10. Evaluate item parameter equivalence across sub-groups	
11. Establish an operational item bank	a. Psychometric results
	b. Clinical input
12. Implement CAT	a. Establish parameters
	b. Simulate across the continuum
13. Create short forms	

continuum. Item parameter equivalence (step 10) is evaluated by examining the existence of DIF across sub-groups (e.g., gender and race/ethnicity). We then carefully reviews analysts' recommendations and clinical input to establish the operational item bank (step11). An item bank at this stage of development is ready for CAT implementation (step 12); test level parameters are proposed (e.g., item selection rules and stopping conditions) and simulated for trait levels across the relevant range of clinical interest. Finally, short forms are created to cover the entire continuum and/or to target specific clinical ranges (step 13).

Testing Options and their Implications for Clinical Adoption

One of the primary advantages of electronic data collection is the capability to provide immediate feedback to the patient and/or their physician. Studies, including some from our group, on the presentation of patient HRQL information have used either graphic or written displays.{Carlson, 2001 9229 /id;Chang, 2002 4491 /id;Detmar, 1998 489 /id;Detmar, 2002 7642 /id;Taenzer, 2000 9245 /id;Velikova, 2001 5360 /id;Velikova, 2002 9247 /id;Velikova, 2004 50014 /id} Functional scales depict performance in physical, role, emotional, and social areas as well as overall quality of life; symptom scales portray common symptoms such as fatigue, nausea, pain, dyspnea, insomnia, appetite loss, and diarrhea. In oncology settings, we{Yount, 2003 21259 /id} and Velikova{Velikova, 2002 9247 /id;Velikova, 2004 50014 /id} have used multiple small graphs to display individual function and symptom profiles over time. We have also explored the use of individual score displays and symptoms on one large graph.{Chang, 2002 4491 /id;Hahn, 2003 9710 /id} Others have focused on written reports that present brief text descriptions of functional issues and symptom profiles, highlighting those that are most problematic to patients.{Carlson, 2001 9229 /id;Detmar, 1998 489 /id;Taenzer, 2000 9245 /id} We and others have found that combined written and graphic reports help identify and prioritize patient problems or concerns, and facilitate communication.{Chang, 2002 4491 /id;Carlson, 2001 9229 /id;Detmar, 1998 489 /id;Higginson, 2001 9239 /id;Taenzer, 2000 9245 /id;Velikova, 2001 5360 /id;Velikova, 2002 9247 /id} Both seem well-received by physicians when they are kept simple and relevant, and are presented immediately.{Buxton, 1998 9228 /id;Carlson, 2001 9229 /id;Detmar, 1998 489 /id;Taenzer, 2000 9245 /id;Velikova, 2001 5360 /id;Velikova, 2002 9247 /id;Velikova, 2004 50014 /id} Graphic presentations may be easier for physicians and patients to grasp.{Velikova, 2004 50014 /id} Immediate presentation of results has been discussed as an important and necessary factor for the incorporating HRQL data to be useful and deemed meaningful in routine clinical practice.{Chang, 2002 4491 /id;Detmar, 1998 489 /id;Taenzer, 2000 9245 /id;Velikova, 2001 5360 /id;Velikova, 2002 9247 /id}

Current software can create computerized instruments that are self-administered by patients on laptop and desktop PCs and over the Internet. We also enable PRO assessment using Personal Data Assistants (PDAs) and Interactive Voice Response (IVR), such that a person can complete fixed-form or CAT assessments from any touch tone telephone. At the core of this system is a data collection platform based on a web-based central data repository. This system can be used to store item text and data assembled from existing datasets, which will serve as the basis for early analyses of items that have already been written and pre-tested by the PRO community. Online

data collection is immediately captured and stored on an item-by-item basis in the data repository. For patients who cannot use online data collection, the resulting data files produced in the course of manual data entry or electronic files contributed directly from alternative data collection efforts can similarly be uploaded into the repository. Finally, data collected using the offline version of the Survey Module are replicated to the central data repository following each assessment or at the end of each day when a research assistant has the opportunity to connect the computer to the Internet.

Working CAT: Simulated and Actual

CAT is seeing increased popularity in information technology (IT) certification, state licensure, college entrance and college placement. To assess minimal competence in software installation and maintenance, Novel and Microsoft initiated programs in the mid-1990's to administer adaptive tests to over 1,000,000 examinees per year. IT certification tests take particular advantage of the strength of CAT to administer unique (and therefore secure) tests, to many people at locations worldwide. CAT is also seeing increased use in state licensure, again related to its capability of frequent, unique-item testing with preserved security. In addition to the pioneering work by the National Council of State Boards of Nursing, programs exist for adaptive administration by the National Association of Securities Dealers, the National Association of Boards of Pharmacy, and the National Board of Medical Examiners. Allied health board certifications using CAT originated by the American Society of Clinical Pathologists are now used by the American Board of Podiatric Surgeons and the American Association of Nurse Anesthetists.

In education, CAT is used for elementary school practice tests, and for college entrance and placement.{Vispoel, 1999}. The college entrance board now administers the Graduate Record Examination (GRE) almost exclusively using CAT, and also administers the Test of English as a Foreign Language (TOEFL) adaptively in locations worldwide where computer access is available. In, 1986 the Psychological Corporation published an adaptive version of the Differential Aptitude Test (DAT), a series of eight tests used since 1947 for placement and vocational guidance in grades 7-12 (McBride, Corpe, & Wing, 1987). CAT is also used to administer their state-wide diagnostic tests. Indiana University Purdue University Indianapolis (IUPUI), uses adaptive tests for both test high school students for potential acceptance, and then again for placement in certain classes. The University Cegeps de Jonquier in Quebec researched an interesting CAT variant which adaptively assesses entering students for placement in English as a Second Language (in the predominantly French speaking program) (Stahl, Bergstrom, & Gershon, 2000). Their exam stands out in two ways – first for their use of hundreds of audio prompts (instead of the typical on-screen text), and second, for their use of CAT and IRT to help identify students who are purposefully cheating on the poor side – attempting to incorrectly answer items that they know the answer to in order to be placed in an easier class.

The most recent new applications in CAT are taking place in rating scale surveys. For example, the Center for Outcomes Research and Education has created item banks for four PROs including Fatigue, Physical Functioning, Pain and Emotional Distress. Adaptive algorithms have been created for use in clinical settings to quickly

ascertain a patient's quality of life in each of these areas and to provide feedback to both the patient and their physician, for immediate consideration in treatment planning. Previous survey instruments were typically judged to be too costly from the perspective of both administration and scoring time, to have immediate clinical utility.{Gershon et al., 2003}

We have developed a CAT algorithm to measure fatigue. We started with fatigue because it is the most common symptom experienced by cancer patients, yet there is currently no standard assessment tool used in clinical practice. CAT expands our options for routine assessment of fatigue (e.g., computer and internet) and allows for easy presentation and interpretation of results. We piloted our fatigue CAT with general cancer patients. IRT analyses of patient responses and discussions with clinicians resulted in a 72-item fatigue item bank. These items were used to create a CAT algorithm that was tested on a sample of outpatients in 3 busy oncology clinics for two assessments (T1 and T2) 2-3 months apart. At T1 all 72 items were administered to all patients via a touch-screen laptop computer. At T2 patients completed all items until the CAT stopped (estimated at 8-10) followed by the remaining FACIT-F items and a 6-item short form. Patients were given the option to complete T2 assessments via touch-screen laptop in the clinic or on the internet. The average CAT assessment speed for the first 224 patients suggests we can conservatively expect 5 questions to be answered per minute, given approximately 10 questions per bank are required for clinical precision, we estimate an average of 2 minutes per bank chosen by the clinical provider for patient assessment.

Field-testing is in progress. We use the item 'I have a lack of energy' as the initial ("screening") item because its response categories showed good distribution across the fatigue continuum, and as an existing FACT-G item, it enabled comparison with different diagnostic groups. A maximum information function is used to select the next item to administer from the item bank.{Gershon, 1995 20986 /id;Parshall, 2001 20978 /id;Gershon, 1995 21298 /id} After each response, the fatigue score and its associated standard error are estimated. We simulate CAT assessments at 100 equidistant trait levels across the range of the trait found in our clinical sampling. For each simulated trait level and item we determine the probability that a given simulee will select a given option. These probabilities serve as weights for a random number generator that selects from among response options in the CAT administration sequence. The process is repeated until a standard error cutoff is obtained or until the maximum number of items is administered. Our simulation studies suggest that accurate assessment of fatigue in individual patients can be obtained with a range of 5-8 items. The only exception is at the ceiling of the measure (very low fatigue) that represents patients who do not usually require additional clinical attention. Our estimation procedure iterates until either the standard error reaches 0.5 or a patient answers 10 items, whichever comes first. We successfully piloted a provisional CAT platform in an oncology clinic.{Davis, 2002 4296 /id} That project enhanced our experience implementing fatigue CAT. We elicited feedback from patients and providers about in-clinic CAT assessment and we assessed the utility and understandability of computer generated graphic reports of fatigue scores. We also monitored the ability of these reports to promote discussions between patients and providers regarding treatment planning and options for on-going care. Feedback from healthcare providers (nurses, physicians and a pharmacist) helped determine their impressions about the utility, understandability and willingness to use fatigue

assessments in routine clinical practice. Both patients (*n*=157) and providers (*n*=22) reported that graphs depicting fatigue scores were understandable and could be useful in clinical practice.{Davis, 2002 4296 /id} Using CAT thereby provides an unprecedented mechanism for brief yet precise, routine symptom monitoring.

Our algorithm considers assessment length. Variable length CAT is typically terminated by a combination of stopping rules: when a specified level of precision (standard error) is reached,{Bergstrom, 1999 21301 /id} when a user-specified maximum number of items has been administered, or when there are no remaining items that will contribute additional information to the estimation of the patient's trait level (as often occurs with a respondent at the ceiling or the floor of the distribution). We will consider enabling a condition to stop testing when a specified level of confidence in a threshold decision is achieved, such as stopping the assessment when the level is sufficiently severe to recommend clinical intervention.{Bergstrom, 1999 21301 /id} Our algorithm handles experimental items by excluding them from scoring but including them in subsequent calibrations. Experimental items may be seeded into the CAT sequence or administered at the end of the assessment; in either case they are not included in the adaptive algorithm or the final patient trait estimate. Our CAT algorithm also considers the importance of calculating scale score transformations when they would be helpful to the end-user. For example, for Rasch measurement models, the typically calculated trait measure range may be -3.5 to +4.0, which is difficult for the end-user to conceptualize. A linear transformation to a 0-100 score or conversion to a *t* distribution is more understandable to both patients and clinicians.{Gershon, 2004 9787 /id}

Thoughts on a Centralized Bank Repository and Cooperative Arrangements for Future Success

While several organizations have already enabled CAT algorithms for PROs {ref CORE; Quality Metric; Walter et al; others} we expect future efforts to generate groundbreaking methods for CAT item selection and person estimation. Our software simulation mode will allow the impact of various CAT parameters to be simulated given any calibrated item bank. The simulation mode "administers" 1-1000 surveys to persons across the trait spectrum, simulating the answers that a person with the given trait level would give, and then choosing the next item based upon the estimation algorithm selected. In this way, the impact of changing IRT models, CAT conditions and bank breadth, depth and quality can all be assessed in detail using provisional calibrations derived from existing datasets and later from network-wide data collection activities.

We envision a national CAT system to include several options for trait estimation, within both 1- and 2-PL applications. The inclusion of multiple trait estimation algorithms will enable researchers to explore the numerous ramifications of selecting a particular algorithm, and further, to examine the potential interactions of each algorithm with other CAT parameters and individual item banks. For example, we are creating a web-based item-banking program that will allow researchers to store items (in multiple languages), their classifications, and resulting statistics. Item templates will allow for several options, including any type of rating scale, integrated graphics and associated sound files.{Bergstrom, 2003 9794 /id} For each item, and for each rating scale option within that item, the item banking system will also serve as a resource for the posting of

statistics associated with each item. These statistics will be stored for each unique analysis and by specific demographic and disease groups. Our survey module considers numerous options for PRO assessment in general, and CAT in particular. Each item is stored in a table along with item parameters and its reference population(s). Numerous test-level parameters are considered as well. While there are many methods of selecting the initial item in CAT, our preferred method is to select an item with response locations across the full range of the trait. Bayesian selection is also commonly used to select the initial item and other models select the first item based on the estimated trait level at a previous assessment. The most significant CAT component to be considered is the formula that calculates the current trait level estimate given the items answered at each point in the assessment. We currently use a maximum likelihood estimation procedure; however, there is not consensus among experts regarding the "best" model to use.

Public Private Partnerships

How will the management and refining of item banks and their applications be sustained over time? We suggest potential models that lend themselves to successful public-private partnerships. Ultimately, we believe on the public side of the partnership the most realistic expectation over time is for continued commitment to fund the best research proposals available on the topic. This is the heart of the NIH mission. The NIH "Roadmap" 5-year commitment to initiating a PRO Measurement Information System ("PROMIS") is an excellent beginning. One way to optimize federal participation beyond this first five years would be through the formation of a PRO Measurement Office or Center within the NIH, whose primary purpose is to stimulate continued new research in support of the many theoretical, technical and practical advances needed to fortify a system in its early stages. Targeted Program Announcements (PA) and Requests for Applications (RFA) would ensure that individual investigators, small businesses, and non-profit organizations recognize the public commitment and consequently choose to direct their research talent in this way.

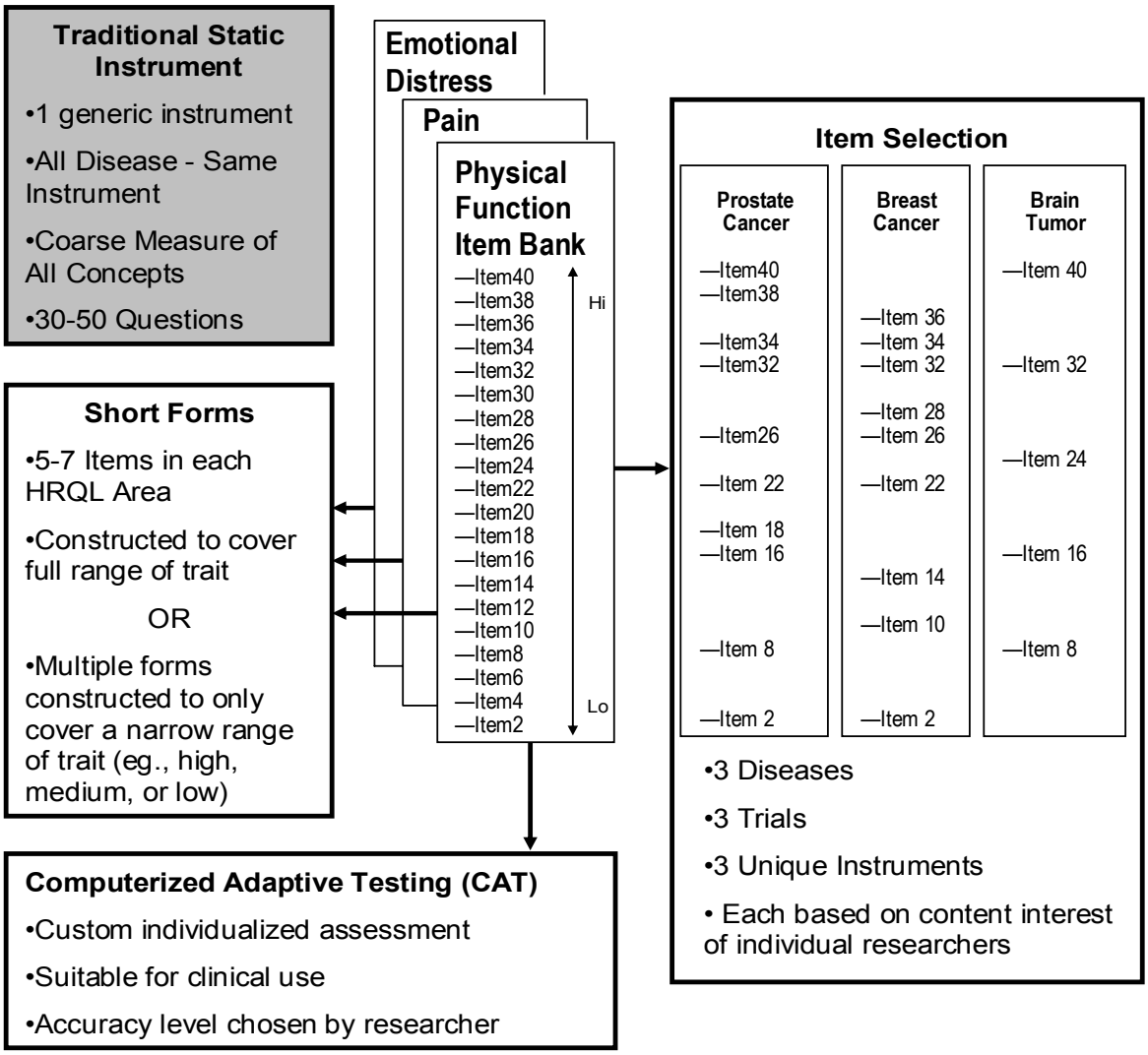
On the "private" side, one possible model could entail a not-for-profit that owns the item bank and related materials. These materials are in turn licensed by the non-profit organization to various organizations, who in turn resell them to others (eg, for-profit testing companies; HMOs, etc.) Using the licensing fees, the non-profit entity may either hire its own staff to manage and develop its banks, or outsource that activity to a contractor. A useful example may be The Iowa Test of Basic Skills (ITBS). ITBS is a series of achievement tests owned by the University of Iowa with content developed by a group of academics at that institution. They license the test for exclusive distribution to Riverside Publishing Company—a division of Houghton Mifflin. Riverside in turn maintains a large sales staff that goes directly to state departments of education, school districts and private schools and "sells" test booklet, answer sheets and scoring services. It is unlikely that a non-profit entity could be created for PRO data that would have sufficient resources in early years to hire a national sales force to promote and sustain the use of the item banks. Similarly, it is unlikely that a large national testing company (such as Riverside), which has a clinical testing division would actually commit sufficient resources in the early years to be considered as an exclusive vendor. It is likely that through a combination of resources, item banks will see maximum use.

A second option to consider is the licensing of IRT/CAT products by a not-for-profit licensing center (“owner”) to for-profit uses by pharmaceutical companies, large provider organizations, or even individual practice groups. Under this arrangement, use by academics would be freely provided and unrestricted. Such a limited distribution approach has worked well for smaller-scale individual PRO questionnaires; however none of these examples has approached the scope envisioned to sustain a common item bank. At minimum, reliable support to minimally maintain and distribute a central electronic resource with application tools is needed. The success of this model will be public support for research and development, as it is unlikely sufficient revenue will be generated, especially in early years, to support necessary measurement and technical advances beyond bank maintenance and distribution of their applications. Perhaps the NIH PRO Measurement Office mentioned earlier could stimulate and support research in support of bank maintenance and improvement.

Conclusion

The use of IRT has many distinct advantages over the classical model of HRQL assessment in practice today. These advantages include: deeper, more comprehensive coverage of each important concept; flexibility in choice of questions used; flexibility in degree of precision desired; availability of multiple short forms; interval measurement contributing to improved statistical power; and capability for individual assessment (real-time clinical monitoring) using CAT. We have focused much of our effort in this area in oncology; however by design these approaches span chronic conditions to the extent that HRQL concepts relevant in oncology (such as pain, fatigue, physical functioning and emotional distress) are common to other conditions. It remains to be seen whether the clinical and clinical research communities will embrace IRT-based PRO assessment applications. We offer some suggestions that might facilitate this.

Figure 1: Calibrated item banks enable users to construct multiple instruments and CAT



Calibrated item banks can be used to easily create a standard Static Instrument. In addition, the banks are used to construct Short Forms, or to enable Computerized Adaptive testing. Researchers and clinicians may also utilize the banks to select items based on unique content interests and formulate custom short-form or full-length instruments. **In every case, using the pre-calibrated item bank allows the short form to draw from the validity of the full bank and produce standardized scores on the same scale.**